DOCUMENT RESUME

ED 434 115 TM 030 076

AUTHOR Wolfe, Edward W.; Moulder, Bradley C.; Myford, Carol M.
TITLE Detecting Differential Rater Functioning over Time (DRIFT) Using a Rasch Multi-Faceted Rating Scale Model.
PUB DATE 1999-00-00
NOTE 41p.; Based on a paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classification; Effect Size; Evaluators; *Item Response Theory; *Rating Scales
IDENTIFIERS *Rasch Model; *Rater Effects

ABSTRACT

This paper describes a class of rater effects that depict rater-by-time interactions. This class of rater effects is referred to as differential rater functioning over time (DRIFT). This article describes several types of DRIFT (primacy/recency, differential centrality/extremism, and practice/fatigue) and Rasch measurement procedures designed to identify these types of DRIFT in rating data. These procedures are applied to simulated data and are shown to be useful in classifying raters as being aberrant or nonaberrant about 95% of the time for primacy, recency, and differential centrality and extremism, particularly for moderate or larger effect sizes. Rates of correct classification for practice and fatigue were lower (about 89%) and statistical power exceeded 0.50 only with very large effect sizes. Type I error rates (i.e., incorrect nomination) were near expected levels in all cases. (Contains 10 tables and 32 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 434 115

Detecting Differential Rater Functioning over Time (DRIFT) Using a Rasch Multi-Faceted Rating Scale Model

Edward W. Wolfe

Bradley C. Moulder

University of Florida

Carol M. Myford

Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Edward Wolfe
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM030076

Author Note

Edward W. Wolfe & Bradley C. Moulder, Foundations of Education; Carol M. Myford, Center for Performance Assessment.

This manuscript is based on a paper that was presented at the annual meeting of the American Educational Research Association in Montreal, Canada, April 1999.

BEST COPY AVAILABLE

Correspondence concerning this article should be addressed to Edward W. Wolfe, College of Education, 1403 Norman Hall, University of Florida, Gainesville, FL 32611. Electronic mail may be sent via Internet to wolfe@nersp.nerdc.ufl.edu.

Abstract

This paper describes a class of rater effects that depict rater-by-time interactions. We refer to this class of rater effects as DRIFT—differential rater functioning over time. This article describes several types of DRIFT (primacy/recency, differential centrality/extremism, and practice/fatigue) and Rasch measurement procedures designed to identify these types of DRIFT in rating data. These procedures are applied to simulated data and are shown to be useful in classifying raters as being aberrant or non-aberrant about 95% of the time for primacy, recency, and differential centrality and extremism, particularly for moderate or larger effect sizes. Rates of correct classification for practice and fatigue were lower (about 89%) and statistical power exceeded .50 only with very large effect sizes. Type I error rates (i.e., incorrect nomination) were near expected levels in all cases.

Detecting Differential Rater Functioning over Time (DRIFT)

Using a Rasch Multi-Faceted Rating Scale Model

Rating scales are commonly used in psychological and educational testing when a systematic procedure is needed for obtaining and reporting the judgments of raters (Linn & Gronlund, 1995). Unfortunately, the use of raters may introduce error into examinee scores for a variety of reasons—unfamiliarity with or inadequate training in the use of the rating scale, fatigue or lapses in attention, deficiencies in some areas of content knowledge that are relevant to making scoring decisions, or personal beliefs that conflict with the values espoused by the scoring rubric. In any case, when raters exhibit problematic rating behaviors, it may be possible to identify unique patterns in the data that correspond to specific types of rater errors. For example, raters who make errors because of fatigue are likely to make more random errors as time progresses. On the other hand, raters who are preoccupied with the goal of assigning ratings that agree with those assigned by other raters are likely to assign a disproportionate number of ratings in the middle of the rating scale.

When the rating task takes place over the period of several hours or several days, concern may arise about the comparability of ratings both between and within raters over time (i.e., "drift"). A number of strategies for preventing or minimizing rater drift have been proposed including initial over training of raters (Kazdin, 1982), providing frequent testing of raters and retraining as necessary (Medley, 1982), periodically recalibrating raters using a previously rated set of stable criterion benchmarks (Johnson & Bolstad, 1973; Kazdin, 1977), providing feedback to raters on their accuracy and levels of interrater agreement (Curran, Beck, Corriveau, & Monti, 1980; DeMaster, Reid, & Twentyman, 1977), delaying all discussions of difficult-to-rate cases until all raters can be involved in those discussions so that all can agree on how to handle such cases when they arise (Reid, 1982), random or surreptitious rescoring of a sample of examinee responses at various points throughout the scoring project to check for evidence of possible drift (Longabaugh,

1982), and employing highly qualified raters to review responses rated by other raters for errors so that incorrectly rated responses can be rescored.

The purpose of this article is to identify several ways that raters may exhibit rater effects that manifest themselves more (or less) apparently as time progresses (i.e., drift) and to describe how these rater effects can be detected using a Rasch multi-faceted rating scale model. The goal of such a practice is to increase the reliability and validity of ratings, a goal that is especially important when ratings are used to make high-stakes decisions (e.g., pass-fail decisions about individual students in educational settings). By identifying when and which raters exhibit drift, scoring leaders can intervene by retraining individual raters or by requiring rescoring of examinee responses that are suspected of being influenced by rater drift. In the following sections, we identify several types of drift, describe a scaling method that can be used to detect these effects, and apply this method to simulated data to verify that the model performs as expected. Finally, we propose further studies of rater drift.

Theoretical Background

Differential Rater Functioning over Time

Previous research has identified several ways that raters may introduce error into examinee scores. A common concern is the extent to which raters' ratings are accurate or are influenced by seemingly random errors (McIntyre, Smith, & Hassett, 1984; Murphy & Balzer, 1989; Sulsky & Balzer, 1988). Another common concern is the extent to which some raters rate systematically more harshly or more leniently than other raters in the rater pool (Engelhard, 1994; Lunz, Wright, & Linacre, 1990). Other concerns include whether raters assign a disproportionate number of ratings in central or extreme scoring categories or whether raters allow their perception of examinees' performances on some assessment tasks to influence their ratings on other assessment tasks (i.e., *halo*) (Engelhard, 1994; Saal, Downey, & Lahey, 1980; Wolfe, Chiu, & Myford, 1999). One class of rater effects that has received relatively little attention in the psychometric literature is rater-by-

time interactions. In this article, we discuss *Differential Rater Functioning over Time* (DRIFT). In the sections that follow we describe how several types of DRIFT manifest themselves in ratings. We also describe a multi-faceted rating scale model that can be used to analyze and detect patterns in rating data that may be indicative of DRIFT.

Primacy/Recency

Primacy and recency effects are well-documented in consumer literature, particularly in the area of taste testing (Berdy, 1969; Dean, 1980; Welch & Swift, 1992; Wolfe & Wolfe, 1997) and to some extent in industrial and organizational psychology (Sayles & Strauss, 1981; Yoder & Staudohar, 1982). These rater effects manifest themselves as increases or decreases in the average rating assigned by a rater as time progresses. In the case of *primacy*, raters assign higher ratings to examinee responses that are evaluated early in the scoring project; that is, raters tend to become more harsh in their ratings as time progresses. The converse is true for *recency* effects (i.e., raters tend to assign higher ratings to examinees that are evaluated later in the project by becoming more lenient as time progresses). In both cases, it is important to note that it is the rater's average rating that changes over time.

Practice/Fatigue

Another common concern in large-scale assessment scoring projects is whether rater training procedures have adequately prepared raters to perform the rating task. If these procedures are not adequate or if raters undergo ongoing training throughout the scoring project (or are periodically recalibrated), then it is possible that the ratings assigned early in the scoring project may not be as accurate as those assigned later in the project. Hence, raters may show higher levels of agreement over time as a result of their gaining *practice* in applying the rating scale to examinee responses. On the other hand, because some rating scales are complex and because many scoring projects last for several days, there may be cause for concern that raters' performance deteriorates over time due to *fatigue*. That is, raters may show lower levels of accuracy as they become tired over

the course of the scoring project. Hence, error is introduced into a rater's ratings over time as a result of both practice and fatigue.

Differential Centrality/Differential Extremism

As described earlier, practice and fatigue manifest themselves as decreases or increases in the accuracy of ratings as time progresses. Another continuum of DRIFT effects describes how rater effects influence the systematic variability of ratings. In the case of ratings assigned during a single time period, we know that some raters exhibit central tendency (i.e., a raters' overuse of the central categories of the rating scale) (Engelhard, 1994; Guilford, 1954; Saal, Downey, & Lahey, 1980). As a scoring project progresses, some raters may use the rating categories in the center of the rating scale more frequently than they did earlier on. We refer to this type of DRIFT as *differential centrality*. The tendency for raters to limit their ratings to the central rating categories after having previously made use of the full range of categories may result from efforts to evaluate raters. For example, raters may realize that they are less likely to be identified as assigning ratings that are in disagreement with other raters if they avoid using the extreme categories of the rating scale. On the other hand, it is possible that raters who are given feedback indicating that they are using the intermediate rating scale categories too frequently may over-correct the situation by assigning too many ratings in the extreme rating scale categories. We refer to this type of DRIFT as *differential extremism* because ratings tend to be assigned more frequently in the extreme rating categories as time progresses. It is important to note that differential centrality and extremism not only introduce error into a rater's ratings as time progresses but that that error systematically decreases or increases (respectively) the variance of the ratings assigned by that rater.

Multi-Faceted Rating Scale Model

The Rasch multi-faceted rating scale model (MFRSM) (Linacre, 1989) describes the probability that an examinee (n) will receive a rating in a particular category (x) by a rater (k) on an assessment task (i). This probability depends on four parameters (Equation 1): the examinee's

proficiency (θ_n), the rater's harshness (λ_k), the task's difficulty (δ_i), and the difficulty of each rating scale threshold (i.e., the threshold between two adjacent rating scale categories, τ_j).

$$P(x|\beta,\lambda,\delta,\tau) = \frac{\exp \sum_{j=0}^{x} \left[\beta_n - \lambda_k - \delta_i - \tau_j\right]}{\sum_{x=0}^{m} \exp \sum_{j=0}^{x} \left[\beta_n - \lambda_k - \delta_i - \tau_j\right]}, \qquad x=0,1,\ldots,m \qquad (1)$$

where, $\tau_0=0$,

and $P(x|\theta,\lambda,\delta,\tau)$ is the probability that examinee *n*'s response to task *i* is assigned a rating of *x* by rater *k* when the rating scale contains *m+1* categories.

Fitting this model to rating data results in a separate parameter estimate and an associated standard error for each examinee, rater, task, and rating scale category in the measurement context. Examinee proficiency depicts the tendency to receive high or low ratings. Rater harshness depicts the rater's tendency to assign high or low ratings, on average. Task difficulty depicts how easy or difficult it is to get high ratings on a specific task. And, rating scale thresholds depict the relative difficulty associated with moving from one category to the next higher category on the rating scale. Of particular interest in the detection of DRIFT are the calibrations for raters (λ). The precision of a rater calibration is depicted by the standard error of that calibration. Rater calibrations are typically scaled so that raters who tend to assign lower ratings to examinees (i.e., *harsh* raters) have higher logit values than raters who assign higher ratings to examinees (i.e., *lenient* raters). Hence, increases and decreases in a rater's calibrations over time would be indicative of primacy and recency effects, respectively.

The method employed for detecting primacy/recency DRIFT in this study allows raters calibrations to "float" from one rating period to the next (Wright, 1996). That is, although examinees and tasks are anchored so that the examinee ability estimates and task difficulty estimates are fixed across time, raters are treated as having a unique calibration for each of the *t*

rating periods. Scaling of data in this manner results in a single calibration for each examinee and each task, but a different calibration for each rater-by-time combination (λ_{rt}). Rater primacy and recency are evidenced by instability of rater calibrations across the rating periods—increases suggesting primacy and decreases implying recency. A significance test can be employed to determine whether increases or decreases in rater calibrations are greater than expected by chance. A standardized difference (Equation 2) is used for this purpose.

$$z = \frac{\lambda_1 - \lambda_2}{\sqrt{\left(SE_{\lambda_1}\right)^2 + \left(SE_{\lambda_2}\right)^2}} \qquad (2)$$

Standardized differences have an expected value of 0.00 and standard deviation of 1.00. Note that this statistic portrays changes in rater calibrations relative to the first rating period (i.e., $t = 1$). Hence, very large standardized differences (e.g., $z > 2.00$) indicate that the rater's harshness has increased at levels beyond that attributable to chance as time progresses (i.e., primacy) while very small standardized differences (e.g., $z < 2.00$) indicate that the rater's harshness has decreased beyond chance levels over time (i.e., recency).

The use of the standardized difference in this manner highlights the fact that the MFRSM portrays a measurement context in which rater harshness calibrations remain invariant across rating periods.[1] Departures in the data from model-generated expected values indicate potentially misfitting raters—misfit that may be indicative of DRIFT effects other than primacy or recency. Patterns of misfitting ratings that are inconsistent with the MFRSM are captured by two statistics associated with each rater calibration. These fit statistics are based on the mean of the squared standardized residuals of the observed ratings from their expected values (Wright & Masters, 1982). The *outfit* statistic is simply an unweighted average of these squared standardized residuals (Equation 3).

$$outfit_{\lambda_x} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{I} z_{nik}^2}{NI} \qquad (3)$$

$$\text{where } z_{nik} = \frac{x_{nik} - E_{nik}}{\sqrt{V(x_{nik})}}$$

$$\text{and } V(x_{nik}) = \sum_{j=0}^{m} (j - E_{nik})P(x \mid \beta, \delta, \lambda, \tau) \quad x=0,1,\ldots,m$$

The *infit* statistic (the weighted mean-square), on the other hand, weights each squared standardized residual by its variance (Equation 4).

$$infit_{\lambda_x} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{I} z_{nik}^2 V(x_{nik})}{\sum_{n=1}^{N}\sum_{i=1}^{I} V(x_{nik})} \qquad (4)$$

Infit and outfit statistics are reported as chi-squares, each divided by its degrees of freedom so that each has an expected value of 1.00 and can range from 0.00 to ∞ (Linacre & Wright, 1994). An infit statistic greater than 1.00 often signals an accumulation of differences between expected and observed ratings near the center of the score distribution, whereas an outfit statistic greater than 1.00 suggests the presence of unexpected residuals in the tails of the score distribution. Fit values less than 1.00 indicate less variability than expected based on the MFRSM. A 0.1 increase in a fit statistic is associated with a 10% increase in unmodeled error beyond that predicted by the model. In general, elements with fit statistic values ranging from 0.8 to 1.4 are considered to show adequate fit to the model (Wright & Linacre, 1994), although the cutoff values tend to vary depending on the purpose for which the ratings are used.

The magnitude of a rater's fit statistic indicates the amount of unmodeled error exhibited by that rater. One can detect changes in the amount of error in that rater's ratings across time by comparing multiple fit statistics, each representing that rater's ratings at a different time point.

Hence, one can detect practice, fatigue, centrality, and extremism as increases or decreases in fit (i.e., unmodeled error) over time. In the context of practice and fatigue, one would expect the rater's fit statistics to decrease or increase (respectively) as time progresses. One would also expect practice and fatigue to have a minimal influence on the variance of the ratings (assuming that the aberrant rater assigns the error-laden ratings according to a distribution that is roughly equivalent in shape to the distributions of ratings assigned by non-aberrant raters). On the other hand, while the fit statistics of raters who commit differential extremism errors should be influenced in ways similar to the fit statistics of raters who commit practice and fatigue errors, the variance of their distributions of ratings should be quite different from the variance of ratings assigned by non-aberrant raters. More specifically, differential extremism should cause the variance of the aberrant raters to be larger than the variance of the ratings of "normal" raters. Interestingly, the introduction of differential centrality into a rater's ratings results in a decrease in rater fit (recall that decreases in infit and outfit indicate decreases in unmodeled error according to the MFRSM) and a decrease in raw score variance.

Fortunately, differences between the pairs of fit statistics for an individual rater can be evaluated statistically, greatly simplifying the diagnosis of primacy, recency, and differential centrality and extremism. Because the fit statistics are reported as chi-square values, each divided by its degrees of freedom, the ratio of two fit indices for a rater approximates an F distribution with $N \times I$ and $N _ I$ (number of examinees times the number of items) degrees of freedom (Hogg & Tanis, 1997) (Equation 5).

$$F = \frac{\frac{\chi_1^2}{df_1}}{\frac{\chi_2^2}{df_2}} \qquad (5)$$

Hence, an *F* value based on two fit statistics that has a null probability less than .05 could be indicative of primacy, recency, or differential centrality or extremism. To further differentiate these types of DRIFT, one can examine a similar *F* ratio that is composed of the variances of raw score ratings assigned by a particular rater at two time periods. Such a ratio of variances also conforms to an *F* distribution with *N* and *N* degrees of freedom (Hayes, 1994) (Equation 6).

$$F = \frac{\sigma_1^2}{\sigma_2^2} \tag{6}$$

As a rule, raters who show increases in both fit and increases in variance across time would be classified as exhibiting differential extremism. Raters who show decrease in fit and decreases in variance across time would be classified as exhibiting differential centrality.[2] Raters who show increases in fit in the absence of changes in variance over time would be classified as exhibiting fatigue. And, raters who show decreases in fit in the absence of changes in variance would be classified as exhibiting the practice effect.

Hence, it seems that statistics associated with the MFRSM could be useful in identifying rater DRIFT. What is unclear is the accuracy of diagnoses of DRIFT that rater evaluators might make based on rater calibrations, fit statistics, and raw score variances. To this end, we carried out a simulation study to identify the decision theoretic power of these procedures for identifying simulated rater aberrance.

Method

Design

Our simulations were designed to evaluate decision rules that would commonly be employed in operational settings for the purpose of detecting DRIFT in ratings data. Our design includes two factors—(a) type of DRIFT effect and (b) size of DRIFT effect. We investigated all six types of DRIFT previously identified in this paper (i.e., primacy, recency, practice, fatigue,

centrality, and extremism). For each of these types of DRIFT, we identified four effect sizes (i.e., *small*, *medium*, *large*, and *very large*). In addition, we generated a *no effect* data set for each set of seed values used. We used this data set a baseline so that we could determine the influence of the introduction of each DRIFT effect _ effect size combination (i.e., we compared all effects over two rating periods by matching each DRIFT effect _ effect size data set with its corresponding *no effect* data set).

The first step in our investigation involved data generation (i.e., simulating ratings so that some raters manifest DRIFT while other raters do not). Next, we performed a MFRSM scaling of each data set. Third, we generated indices based on the MFRSM analyses for identifying each type of DRIFT under investigation. Finally, we applied predetermined criteria to each simulated rater's calibration, fit index, or raw score standard deviation and "nominated" raters as exhibiting DRIFT based on predetermined decision rules. We performed decision theoretic analyses on these nominations by comparing nomination status (i.e., DRIFT versus non-DRIFT) to true status as designated during data generation. The following sections elaborate on each step of these procedures.

Data Generation

By crossing the six levels of DRIFT effects with the five levels of effect size, we specified a total of 30 cells in our design. For each cell, we generated 30 replications resulting in a total of 900 simulated data sets. We simulated each data set to portray a measurement situation in which each of 100 examinees responded to a single task, and 100 raters rated each examinee response using a five-point rating scale. We generated data based on procedures suggested by Harwell, Stone, Hsu, and Krisci (1996). That is, first, we specified an item response model for data generation. For our purposes, a two-parameter logistic-rater model was appropriate (Equation 7). This model (Wolfe, 1998) contains five parameters—θ_n represents the examinee's ability (location), δ_i represents the assessment task's difficulty (location), τ_j represents the rating scale threshold difficulty (location),

λ_k represents the rater's harshness (location), and γ_k represents the rater's centrality (slope). For all data sets, we sampled examinee ability from a N(0,1) distribution, we set task difficulty to 0.00 for the single assessment task, and we set the four rating scale step difficulties to –2.00, -1.00, 1.00, and 2.00.

$$z = \frac{\lambda_1 - \lambda_2}{\sqrt{\left(SE_{\lambda_1}\right)^2 + \left(SE_{\lambda_2}\right)^2}} \qquad (7)$$

We introduced rater effects by altering λ_k and γ_k for some raters (to introduce primacy/recency and centrality/extremism into raters' ratings, respectively). In addition, some raters' ratings by randomly generated values that maintained the same distributional shape of the raw ratings (hereafter referred to as *% Random*) in order to introduce practice/fatigue effects. In each data set, we designated 90 of the 100 raters (90%) as non-DRIFT raters (i.e., they were not modeled to manifest any DRIFT in their ratings). For these non-DRIFT raters, λ_k equaled 0.00, γ_k equaled 1.00, and *% Random* equaled 0. For the 10 raters in each data set we designated as manifesting DRIFT, we determined the type and size of the DRIFT by the cell in the design matrix, and we introduced this DRIFT into the rater's ratings by altering the appropriate rating scale parameter as indicated by the rules shown in Table 1. More specifically, primacy and recency raters had the λ_k values added into Equation 7 as indicated in the first two rows of Table 1 (i.e., an increase of 0.50 logits per effect size level). Differential centrality and extremism raters had the γ_k values as shown in rows three and four multiplied into Equation 7 as indicated by effect size. Practice and fatigue raters had the percent of ratings indicated by the effect size level altered by replacing θ_n with a value generated from an orthogonal (i.e., uncorrelated) ability distribution that we sampled for every rater-by-examinee combination.

==========

INSERT TABLE 1 ABOUT HERE

==========

By comparing each *no effect* data set with each of the remaining effect size data sets, we were able to determine the degree to which the relevant DRIFT indices were influenced by the introduction of modeled DRIFT. We generated these corresponding data sets using identical same seed values, and the seed values differed between the 30 replications within a cell. As shown in Table 1, we generated all *no effect* data sets by setting λ_k equal to 0.00, γ_k equal to 1.00, and % *random* equal to 0. For primacy, recency, differential centrality and extremism, and fatigue, the *no effect* data sets served as the *Time 1* data, and the corresponding data sets associated with each effect size level served as the relevant *Time 2* data. For practice, the only effect type in which the DRIFT effect disappears over time, the data sets associated with each effect size level served as the *Time 1* data, and the corresponding *no effect* data set served as the *Time 2* data.

Scaling

We simultaneously scaled each pair of data sets (i.e., the *Time 1* and corresponding *Time 2* data sets as described in the previous subsection) to the MFRSM (Equation 1) using *Facets* (Linacre, 1997). We treated raters as unique elements of the measurement context at each time index by assigning a different rater identifier for *Time 1* and *Time 2*. By treating raters as being unique at the two rating periods, we were able to produce a pair of rater calibrations, both calibrated to the same underlying scale, for each rater. These pairs of rater calibrations and the associated fit statistics (along with the raw score variances for these raters) served as the indices for identifying DRIFT as described in the next subsection.

Nomination Criteria

The criteria we used to diagnose DRIFT in the simulated data are shown in Table 2. We used the standardized difference between the *Time 1* and *Time 2* rater calibrations to diagnose rater primacy and recency. For differential centrality and extremism, there are three indices that provide

relevant information (all being *F* ratios—these ratios based on changes in rater infit, outfit, and raw score variance between *Time 1* and *Time 2*). For practice and fatigue effects, we used *F* ratios based on rater infit and outfit to diagnose DRIFT. For each index, if a rater's index met the criteria shown in Table 1, we nominated the rater as exhibiting DRIFT. Otherwise, we nominated the rater as exhibiting no DRIFT. We compared nomination status to true status (as designated during data generation), and we then performed decision theoretic analyses using 2 _ 2 tables to examine the proportion of raters who were correctly and incorrectly diagnosed for each of the three DRIFT continua. We averaged proportions of correct and incorrect decisions across the 30 replications within each DRIFT effect _ effect size combination. This resulted in 4 tables each for primacy and recency (1 index _ 4 effect sizes), 12 tables each for differential centrality and extremism (3 indices × 4 effect sizes), and 8 tables each for fatigue and practice (2 indices _ 4 effect sizes).

==========

INSERT TABLE 2 ABOUT HERE

==========

Results

The results of our simulations indicate that the raw score distributions, rater calibrations, and rater fit statistics behave in predictable ways when we introduced the various types and magnitudes of DRIFT into the simulated ratings. More specifically, the introduction of primacy decreases the average rating assigned by a rater while increasing the average rater calibration, and the introduction of recency has the opposite effect. The introduction of differential centrality decreases the raw score variance and the fit statistics, while the introduction of differential extremism increases these two statistics. Of course, the introduction of centrality and extremism also decreases and increases the raw score variances, respectively. On the other hand, the introduction of practice and fatigue effects has a minimal impact on the raw score distributions. But, as one would expect, practice and fatigue manifest themselves as decreases and increases, respectively, in rater fit.

Our simulations also show that the criteria we outlined previously in this paper can be used to detect most types of rater DRIFT. The standardized difference seems to be an excellent index for detecting changes in rater calibrations across time (i.e., primacy and recency). This statistic correctly nominated most DRIFT raters when the effect size was 1.00 logit in size with minimal incorrect nomination of non-DRIFT raters. For differential centrality and extremism, F_{infit} minimized incorrect nominations and maximized correct nominations for differential extremism. F_{outfit} increased the rate of incorrect nominations slightly for differential centrality, but also maximized the rate of correct nomination of DRIFT raters. Interestingly, fit statistics provided better diagnosis of differential centrality and extremism than did raw score variances. Most DRIFT raters were correctly identified when the variance of DRIFT ratings was reduced to 50% of the true variance (differential centrality) and when the variance of DRIFT ratings was 1.25 times greater than the true variance (differential extremism). The F_{infit} and F_{outfit} statistics were inadequate for detecting practice and fatigue effects at all effect size levels.

Raw Scores

Table 3 shows the raw score descriptive statistics for all raters and for aberrant raters for each DRIFT effect _ effect size combination. These figures demonstrate that the introduction of DRIFT into the ratings of some raters influenced the mean, standard deviation, skewness, and kurtosis of the distributions of ratings in predictable ways. The first two rows of data show that the introduction of increasingly larger primacy effects consistently lowered the average rating. This change in the mean of the ratings caused the distribution to become more positively skewed and more leptokurtic. Of course, the introduction of recency influenced the distribution of ratings in the opposite direction but with a similar magnitude. Also, as would be expected, the introduction of differential centrality caused the ratings to be more leptokurtic, resulting in a significant shrinkage of the standard deviation across increases in effect size. The introduction of differential centrality also resulted in no change in the mean or skewness of the distribution of ratings. Again, the contrary DRIFT effect, differential extremism, had the opposite effect, increasing the standard

deviation and causing the distribution to become platykurtic. Finally, the introduction of practice and fatigue effects had only a minimal influence on the shape of the distribution with the most noticeable effect being a flattening of the distribution (i.e., shrinking kurtosis) with increasingly larger effect sizes.

==========

INSERT TABLE 3 ABOUT HERE

==========

Calibrations and Fit Indices

Table 4 shows the average values of the rater calibrations, standard errors of those calibrations, and infit and outfit statistics for all raters and for the DRIFT raters for each DRIFT effect _ effect size combination. As shown by these figures, the rater calibrations of the DRIFT raters under the primacy and recency conditions increase and decrease, respectively, with progressively larger effect sizes. And, as one would expect, the similar effect sizes for the primacy and recency effects result in similar increases and decreases in the average DRIFT rater calibrations. Under the differential centrality and extremism conditions, the rater infit and outfit decrease and increase, respectively. That is, differential centrality results in smaller residuals from the MFRSM expected values while differential extremism results in larger residuals from the MFRSM expected values. And, as was shown in Table 3, differential centrality and extremism also resulted in decreases and increases, respectively, of raw score standard deviations with progressively larger effect sizes. The introduction of fatigue, as shown in the last row of Table 4, resulted in increased rater fit statistics, but, according to Table 3, did not result in large increases in the raw score standard deviations. Also, recall that the *no effect* effect size for the practice effect was the second time period in our simulations, and each of the remaining effect size levels was designated as the first time period. Therefore, the data in Table 4 suggest that the practice effect resulted in large decreases in rater fit (i.e., decreases in unmodeled error) but did not, according to Table 3, result in large decreases in raw score standard deviations. Finally, note that the summary statistics

for the *no effect* condition vary slightly across effect types. This is because we calibrated the data from the *no effect* and the cooresponding effect sizes simultaneously, which resulted in some of the DRIFT effects being absorbed by the *no effect* parameter estimates.

==========

INSERT TABLE 4 ABOUT HERE

==========

Decision Theoretic Analysis

Given the fact that the raw score and rater calibration summaries behaved in way that one would predict based on the type and size of DRIFT introduced into the data, the next question is, "Can these statistics be used to identify DRIFT using objective statistical criteria." The remaining three subsections address this question by the presenting the results of the decision theoretic analyses of the nomination criteria described previously.

Note that Tables 5-10 compare the percentages of raters who were correctly and incorrectly nominated as exhibiting each DRIFT effect under each effect size condition. For example, the first data row of Table 5 shows the percent of non-DRIFT raters who were correctly identified using our nomination criteria (Table 2) for the primacy and recency effects. Recall that we designated 90% of the raters as being non-DRIFT raters so perfect nomination for the *small* primacy effect would result in 90% of the raters falling in the first data row of the first data column of Table 5. The fact that the observed number is 85.5% indicates that 4.5% of the non-DRIFT raters were incorrectly nominated as exhibiting primacy as shown in the second data row of the first data column of Table 5(i.e., a Type I error). The third data row shows that 5.85% of the DRIFT raters were not identified as exhibiting primacy (i.e., a Type II error), and the fourth data row shows that 4.17% of the DRIFT raters were correctly nominated as primacy raters when a *small* effect size was modeled.

Primacy/Recency

Recall that for primacy and recency detection, a rater was nominated for exhibiting DRIFT if the standardized difference of the two calibrations was larger than +2.00 (recency) or smaller

than –2.00 (primacy). Table 5 shows that, overall, the standardized difference was an excellent indicator of these two types of DRIFT. Generally, there were few incorrect nominations based on this statistic (as evidenced by the low percentage of raters falling into the second and sixth data rows of this table). Note that these values are close to (and in most cases slightly smaller than) the expected Type I error rate associated with an absolute z value of 2.00 (4.55%). As shown by the pairs of rows associated with the DRIFT raters for each of these types of DRIFT, as the effect size increased, the percent of correct identifications increased and the percent of non-nominated DRIFT raters decreased. For both primacy and recency, the standardized difference accurately identified most of the DRIFT raters when the effect size was *medium* (i.e., an increase or decrease of one logit in rater harshness).

==========

INSERT TABLE 5 ABOUT HERE

==========

Differential Centrality/Extremism

Table 6 shows the accuracy of nominations for differential centrality and extremism based on the F_{infit} statistic. Recall that this statistic is the ratio of $infit_{j,k}$ from two rating periods for a single rater, with large values indicating that the fit statistics are statistically different. For differential centrality, the numerator of F_{infit} is the $infit_{j}$ from the first rating period, and the denominator is the $infit_{j}$ for the second rating period. The opposite is true for differential extremism. As shown in Table 6 (data rows 2 and 6), the F_{infit} statistic resulted in a fairly small portion of the non-DRIFT raters being incorrectly nominated for DRIFT across the effect size levels (about 2.5% for differential centrality and about 3.5% for differential extremism—both are smaller than the expected Type I error rate of 5%). In addition, as the effect size increased, so did the percent of correctly identified DRIFT raters (see data row 4 of Table 6). Note that most of the DRIFT raters (i.e., 7.5%) were correctly identified for differential centrality when the effect size was moderate (i.e., when the variance of the DRIFT rater's ratings was about 50% of that of the non-DRIFT rater), but that

extremism was detected for most of the DRIFT raters (7.03%) when the effect size was small (i.e., the variance of the DRIFT rater's ratings was about 125% of the size of the variance of the non-DRIFT rater). Recall from Table 1 that these values are not symmetrical across effect sizes with respect to effect type because the effect sizes for differential centrality and extremism were not proportional.

==========

INSERT TABLE 6 ABOUT HERE

==========

Table 7 shows the accuracy of the differential centrality and extremism nominations that were based on the F_{outfit} statistic. The F_{outfit} statistic is formed in the same way as the F_{infit} statistic, except the former is based on a rater's $outfit_{jk}$. Note that the accuracy of DRIFT nomination for the outfit statistic is similar to that observed using the infit statistic. However, it should be noted that the outfit statistic resulted in a larger number of incorrect nominations of non-DRIFT raters as exhibiting DRIFT (still, at a rate about equal to the expected Type I error rate of 5%). In addition, for differential centrality (but not for differential extremism), the outfit statistic resulted in slightly better correct identification of DRIFT raters than was the case for the infit statistic.

==========

INSERT TABLE 7 ABOUT HERE

==========

Table 8 shows the percent of correct and incorrect DRIFT nominations for differential centrality and extremism based on the $F_{variance}$ statistic. Recall that differential centrality and extremism should result in decreases and increases in the variance of a rater's ratings over time, respectively. Also recall that, for differential centrality, the numerator of $F_{variance}$ is the variance of the rater's ratings from the first rating period, and the denominator is the variance of the rater's ratings from the second rating period. The opposite is true for differential extremism. Table 8 shows that the raw score variance resulted in similar rates of correct and incorrect identification of DRIFT

raters as was observed using the $infit_{zk}$ statistic, but that the raw score variances were not quite as accurate (i.e., compare data rows 4 and 8 of Tables 6 and 8).

==========

INSERT TABLE 8 ABOUT HERE

==========

Fatigue/Practice

Table 9 shows the accuracy rates for the nomination of raters under the practice and fatigue conditions using the F_{infit} statistic. These values show that the rate of incorrect nomination of non-DRIFT raters was low (i.e., about 3.5%--lower than the expected Type I error rate of 5%) across all effect size levels, and the correct nomination of DRIFT raters increased as the effect size increased. But even with a *very large* effect size (i.e., 75% of the ratings during the aberrant time period were random), only about half of the DRIFT raters were correctly nominated.

==========

INSERT TABLE 9 ABOUT HERE

==========

Table 10 shows the corresponding percentages for practice and fatigue effects based on the F_{outfit} statistic. As shown, the outfit statistic resulted in a larger percent of incorrectly nominated non-DRIFT raters (about 5% across the effect size levels) than was the case for the F_{infit} statistic (i.e., compare data rows 2 and 6 of Tables 9 and 10). However, the outfit statistic resulted in slightly better identification of true DRIFT raters (i.e., compare data rows 4 and 8 of the two tables). Still, many DRIFT raters were not nominated (i.e., data rows 3 and 7), even when the effect size was very large.

==========

INSERT TABLE 10 ABOUT HERE

==========

Discussion

Overall, our data suggest that the detection methods that we explored in this paper may be very useful for identifying DRIFT. We provided evidence that our simulated data conform to our expectations by demonstrating that raw score distributions and rater calibrations and fit statistics are influenced in ways one would expect by the introduction of various types and magnitudes of DRIFT. In addition, we have shown that, with the exception of the practice and fatigue effects, the MFRSM-based criteria we explored result in low levels of misdiagnosis of DRIFT (i.e., Type I statistical errors) while providing correct diagnosis of most simulated DRIFT raters (i.e., minimizing Type II statistical errors). In addition, these accuracy rates were accomplished with modest effect sizes. What is particularly useful about the demonstrated application of these methods as they pertain to the MFRSM is the fact that the rater calibrations and fit statistics proved to be considerably more useful than did the raw scores. Not only did the MFRSM-based criteria allow us to more accurately identify some types of DRIFT than was possible using the raw score equivalent (e.g., differential centrality and extremism), but these criteria also allowed us to have at least limited success in detecting DRIFT effects for which there is no raw score method available (e.g., practice and fatigue).

We believe that these DRIFT detection methods would be very useful in large-scale rating projects such as those typical of statewide standards-based performance assessments and national or international performance-based assessments (e.g., NAEP and TIMSS performance tasks, the SAT written section, or TOEFL writing and listening assessments). In contexts such as these, DRIFT analyses could routinely be performed, and the resulting statistics could be monitored so that raters exhibiting DRIFT could be identified before a large numbers of responses needed to be rerated or error-laden ratings are reported. These raters could be retrained or recalibrated prior to

assigning additional ratings that would be reported to examinees. As a result, considerable cost could be saved in the rating project while maintaining high levels of reliability. For some rating projects, particularly those that employ image-based response documentation so that responses can be distributed and ratings can be documented electronically, these procedures could be automated so that raters are evaluated and provided with feedback in real time.

Of course, the major limitation of our work is the fact that it is based only on simulations. As a result, we may be working with data that is "cleaner" than would be encountered in most operational settings. Specifically, we introduced only one level and one type of DRIFT in each cell of our design. It is likely that there are several types of DRIFT in any single operational data set, and almost certainly each of those types of DRIFT is represented by a continuum of effect sizes across the rater pool. It is unclear how the interaction of various DRIFT effects and the ranges of effect sizes would influence the accuracy of the diagnoses made in our simulated data.

Hence, an important direction for future research concerning these DRIFT detection methods is to apply them in operational settings. Specifically, we would like to apply these methods to a variety of operational data sets to determine the types of DRIFT that seem to be apparent in these data sets and the ranges of effect sizes that they contain. In addition, studies with a more qualitative flavor should be performed to determine the conditions under which various types of DRIFT are likely to occur and the optimal time sampling strategies that should be used for detecting those types of DRIFT. From a more quantitative orientation, additional simulation studies would be useful for performing a more fine-grained analysis of the Type I and Type II error rates of more closely spaced effect sizes than we investigated and for determining how the alteration of selection rates (i.e., Type I error rate) influences the power of the DRIFT detection indices. In addition, more complex rating situations should be simulated to determine whether the levels of accuracy observed in this study are maintained when different raters exhibit a variety of types of DRIFT simultaneously. Of course, there are a variety of fit indices that have been formulated for

determining model-data fit, and these should also be examined to determine if any of them are more sensitive than the infit and outfit statistics examined in this study.

Endnotes

[1] The MFRSM also specifies that a common rating scale structure applies to each task (i.e., that τ_i is constant across tasks). It should be noted that it is possible to define a multi-faceted model that allows the rating scale structure to vary from one task to another (or even from one rater to another). We refer to this model as a multi-faceted partial credit model (MFPCM). Although our examples use only the MFRSM, the methods we describe in this article can be used with the MFPCM as well.

[2] Note that this is not always the case. In the unlikely event that the raw score distribution is platykurtic and the rating scale threshold difficulties are widely spaced, centrality might actually increase rater fit (Wolfe, Chiu, & Myford, 1999).

References

Berdy, D. (1969). Order effects in taste tests. *Journal of the Market Research Society, 11*, 361-371.

Curran, J.P., Beck, J.G., Corriveau, D.P., & Monti, P.M. (1980). Recalibration of raters to criterion: A methodological note for social skills research. *Behavioral Assessment, 2*, 261-268.

Dean, M.L. (1980). Presentation order effects in product taste tests. *Journal of Psychology, 105*, 107-110.

DeMaster, B., Reid, J., & Twentyman, C. (1977). The effects of different amounts of feedback on observer's reliability. *Behavior Therapy, 8*, 317-329.

Engelhard, G.J. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Guilford, J. P. (1954). *Psychometric Methods* (2nd ed.). New York, NY: McGraw-Hill.

Harwell, M., Stone, C.A., Hsu, T.C., & Krisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Hayes, W.L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.

Hogg, R.V., & Tanis, E.A. (1997). *Probability and statistical inference* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Johnson, S.M., & Bolstad, O.D. (1973). Methodological issues in naturalistic observation: Some problems and solutions for field research. In L.A. Hamerlyunck, L.C. Handy, & E.J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice*. Champaign, IL: Research Press.

Kazdin, A.E. (1982). Observer effects: Reactivity of direct observation. In D.P. Hartmann (Ed.), *Using observers to study behavior* (pp. 5-19). San Francisco: Jossey-Bass.

Kazdin, A.E. (1977). Artifact, bias, and complexity: The ABC's of reliability. *Journal of Applied Behavior Analysis, 10*, 141-150.

Linacre, J.M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA.

Linacre, J.M., & Wright, B.D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, *8*, 360-361.

Linn, R.L., & Gronlund, N.E. (1995). *Measurement and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Merrill.

Longabaugh, R. (1982). The systematic observation of behavior in naturalistic settings. In H.C. Triandis & J.W. Berry (Eds.), *Handbook of cross-cultural psychology. Volume 2: Methodology* (pp. 57-126). Boston: Allyn and Bacon.

Lunz, M.E., Wright, B.D., & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*, 331-345.

McIntyre, R.M., Smith, D.E., & Hassett, C.E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, *69*, 147-156.

Medley, D.M. (1982). Systematic observation. In H.E. Mitzel (Ed.), *Encyclopedia of educational research. Volume 4* (5th ed., pp. 1841-1851). New York: Macmillan.

Murphy, K.R., & Balzer, W.K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619-624.

Reid, J.B. (1982). Observer training in naturalistic research. In D.P. Hartmann (Ed.), *Using observers to study behavior* (pp. 37-50). San Francisco: Jossey-Bass.

Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413-428.

Sayles, L.R., & Strauss, G. (1981). *Managing human resources* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Sulsky, L.M., & Balzer, W.K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*, 497-506.

Welch, J.L., & Swift, C.O. (1992). Question order effects in taste testing of beverages. *Journal of the Academy of Marketing Science, 20*, 265-268.

Wolfe, E.W. (1998). *A two-parameter logistic rater model (2PLRM): Detecting rater harshness and centrality*. Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Wolfe, E.W., & Chiu, C.W.T., Myford, C.M. (1999). *Detecting rater effects with a multi-faceted rating scale model* (Report 97-02). Princeton, NJ: ETS.

Wolfe, E.W., & Wolfe, C.L. (1997). Questioning order in the court of beer judging—A study of the effects of presentation order in beer competitions. *Brewing Techniques, 5*(2), 44-49.

Wright, B.D., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Wright, B.D. (1996). Comparisons require stability. Rasch Measurement Transactions, 10, 506.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Yoder, D., & Staudohar, P.D. (1982). *Personnel management and industrial relations* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Table 1

Data Generation Parameter Specifications

Effect	Parameter	Effect Size				
		None	Small	Medium	Large	Very Large
Primacy	λ	0.00	0.50	1.00	1.50	2.00
Recency	λ	0.00	-0.50	-1.00	-1.50	-2.00
Centrality	γ	1.00	1.25	1.50	1.75	2.00
Extremism	γ	1.00	0.67	0.50	0.40	0.33
Practice	*% Random*	0	10	25	50	75
Fatigue	*% Random*	0	10	25	50	75

Table 2

DRIFT Nomination Criteria

Effect Type	Index			
	$z_{\lambda k}$	F_{infit}	F_{outfit}	F_{variance}
Primacy	$z_{\lambda k} < -2.00$	NA	NA	NA
Recency	$z_{\lambda k} > 2.00$	NA	NA	NA
Centrality	NA	$P(F_{\text{infit}}) < .05$	$P(F_{\text{outfit}}) < .05$	$P(F_{\text{variance}}) < .05$
Extremism	NA	$P(F_{\text{infit}}) < .05$	$P(F_{\text{outfit}}) < .05$	$P(F_{\text{variance}}) < .05$
Practice	NA	$P(F_{\text{infit}}) < .05$	$P(F_{\text{outfit}}) < .05$	NA
Fatigue	NA	$P(F_{\text{infit}}) < .05$	$P(F_{\text{outfit}}) < .05$	NA

Note: $z_{\lambda k}$ represents the standardized difference of two λ_k estimates (Equation 2). For primacy, we expect $\hat{\lambda}_{k2} - \lambda_{k1}$ to be negative, and for recency we expect $\lambda_{k2} - \lambda_{k1}$ to be positive. F_{infit} and F_{outfit} represents the ratio of two λ_k infit and outfit statistics, respectively (Equation 5). These values should be large for extremism and fatigue when $fit_{\lambda k2}$ is divided by $fit_{\lambda k1}$ and should be large for centrality and practice when $fit_{\lambda k1}$ is divided by $fit_{\lambda k2}$. F_{variance} represents the ratio of two raw score variances for a particular rater (Equation 6). These values should be large for centrality when $V(x_{k1})$ is divided by $V(x_{k2})$ and should be large for extremism when $V(x_{k2})$ is divided by $V(x_{k1})$. *NA* indicates that the index in question is not relevant for detecting the DRIFT effect.

Table 3

Raw Score Summary Statistics for All Raters and DRIFT Raters

Effect	Group	Statistic	Effect Size				
			None	Small	Medium	Large	Very Large
Primacy	All raters	Mean	1.84	1.82	1.79	1.76	1.74
		SD	1.16	1.17	1.17	1.19	1.19
		Skew	0.29	0.30	0.31	0.32	0.32
		Kurtosis	-0.17	-0.20	-0.21	-00.25	-0.26
	DRIFT	Mean	1.84	1.59	1.33	1.07	0.85
		SD	1.16	1.15	1.12	1.07	0.99
		Skew	0.29	0.38	0.52	0.73	0.94
		Kurtosis	-0.16	-0.13	-0.07	0.09	0.36
Recency	All raters	Mean	1.84	1.87	1.89	1.92	1.94
		SD	1.16	1.17	1.18	1.19	1.21
		Skew	0.29	0.28	0.27	0.26	0.25
		Kurtosis	-0.17	-0.23	-0.27	-0.34	-0.40
	DRIFT	Mean	1.84	2.11	2.38	2.62	2.89
		SD	1.16	1.19	1.20	1.19	1.15
		Skew	0.29	0.20	0.05	-0.12	-0.38
		Kurtosis	-0.16	-0.44	-0.74	-0.95	-1.03
Centrality	All raters	Mean	1.84	1.84	1.84	1.84	1.84
		SD	1.16	1.16	1.15	1.14	1.13
		Skew	0.29	0.30	0.29	0.29	0.30
		Kurtosis	-0.17	-0.15	-0.10	-0.09	-0.05
	DRIFT	Mean	1.84	1.84	1.83	1.83	1.83
		SD	1.16	1.05	0.95	0.87	0.81
		Skew	0.29	0.32	0.31	0.28	0.27
		Kurtosis	-0.16	0.40	1.02	1.63	2.23
Extremism	All raters	Mean	1.84	1.85	1.85	1.85	1.85
		SD	1.16	1.19	1.21	1.22	1.22
		Skew	0.29	0.29	0.28	0.28	0.28
		Kurtosis	-0.17	-0.31	-0.37	-0.42	-0.44
	DRIFT	Mean	1.84	1.88	1.90	1.88	1.92
		SD	1.16	1.39	1.53	1.62	1.68
		Skew	0.29	0.22	0.18	0.17	0.13
		Kurtosis	-0.16	-0.99	-1.32	-1.48	-1.59
Practice	All raters	Mean	1.84	1.84	1.84	1.84	1.84
		SD	1.16	1.17	1.16	1.16	1.16
		Skew	0.29	0.29	0.29	0.29	0.30
		Kurtosis	-0.17	-0.21	-0.19	-0.19	-0.19
	DRIFT	Mean	1.84	1.85	1.85	1.84	1.85
		SD	1.16	1.19	1.17	1.17	1.17
		Skew	0.29	0.29	0.29	0.29	0.29
		Kurtosis	-0.16	-0.30	-0.21	-0.20	-0.20
Fatigue	All raters	Mean	1.84	1.84	1.84	1.84	1.84
		SD	1.16	1.17	1.17	1.16	1.16
		Skew	0.29	0.29	0.29	0.29	0.30
		Kurtosis	-0.17	-0.20	-0.19	-0.19	-0.19

	DRIFT	Mean	1.84	1.84	1.85	1.85	1.86
		SD	1.16	1.18	1.17	1.17	1.16
		Skew	0.29	0.29	0.29	0.29	0.29
		Kurtosis	-0.16	-0.26	-0.23	-0.22	-0.20

Table 4

Average Rater Calibrations and Associated Statistics for All Raters and DRIFT Raters

Effect	Group	Statistic	Effect Size				
			None	Small	Medium	Large	Very Large
Primacy	All raters	λ_k	-0.05	-0.02	0.01	0.04	0.07
		$SE_{\lambda k}$	0.12	0.12	0.12	0.12	0.12
		$Infit_{\lambda k}$	0.99	1.00	1.00	1.01	1.00
		$Outfit_{\lambda k}$	1.01	1.02	1.02	1.03	1.02
	DRIFT	λ_k	-0.06	0.25	0.55	0.87	1.16
		$SE_{\lambda k}$	0.12	0.12	0.12	0.12	0.12
		$Infit_{\lambda k}$	0.99	0.99	1.00	1.02	1.01
		$Outfit_{\lambda k}$	1.01	1.00	1.01	1.02	1.02
Recency	All raters	λ_k	-0.05	-0.09	-0.12	-0.16	-0.20
		$SE_{\lambda k}$	0.12	0.12	0.12	0.13	0.13
		$Infit_{\lambda k}$	0.99	1.00	1.01	1.02	1.01
		$Outfit_{\lambda k}$	1.01	1.02	1.03	1.03	1.03
	DRIFT	λ_k	-0.06	-0.38	-0.73	-1.07	-1.51
		$SE_{\lambda k}$	0.12	0.13	0.14	0.15	0.17
		$Infit_{\lambda k}$	0.99	1.03	1.10	1.13	1.17
		$Outfit_{\lambda k}$	1.00	1.05	1.11	1.12	1.15
Centrality	All raters	λ_k	-0.05	-0.06	-0.06	-0.06	-0.06
		$SE_{\lambda k}$	0.13	0.13	0.13	0.13	0.13
		$Infit_{\lambda k}$	1.02	1.01	1.00	0.99	0.97
		$Outfit_{\lambda k}$	1.05	1.03	1.02	1.01	0.99
	DRIFT	λ_k	-0.06	-0.10	-0.13	-0.15	-0.17
		$SE_{\lambda k}$	0.13	0.13	0.13	0.13	0.13
		$Infit_{\lambda k}$	1.02	0.80	0.64	0.52	0.44
		$Outfit_{\lambda k}$	1.04	0.80	0.64	0.51	0.42
Extremism	All raters	λ_k	-0.05	-0.05	-0.04	-0.03	-0.03
		$SE_{\lambda k}$	0.12	0.12	0.12	0.12	0.12
		$Infit_{\lambda k}$	0.92	0.98	1.01	1.04	1.04
		$Outfit_{\lambda k}$	0.93	0.99	1.03	1.06	1.07
	DRIFT	λ_k	-0.06	0.01	0.05	0.11	0.10
		$SE_{\lambda k}$	0.12	0.12	0.12	0.12	0.12
		$Infit_{\lambda k}$	0.92	1.39	1.74	1.96	2.13
		$Outfit_{\lambda k}$	0.93	1.44	1.81	2.05	2.24
Practice	All raters	λ_k	-0.04	-0.04	-0.05	-0.05	-0.04
		$SE_{\lambda k}$	0.12	0.12	0.12	0.12	0.12
		$Infit_{\lambda k}$	0.98	0.99	1.00	1.01	1.03
		$Outfit_{\lambda k}$	0.99	1.00	1.02	1.03	1.05
	DRIFT	λ_k	-0.05	-0.05	-0.05	-0.06	-0.05
		$SE_{\lambda k}$	0.12	0.12	0.12	0.12	0.12
		$Infit_{\lambda k}$	0.98	1.02	1.12	1.22	1.37
		$Outfit_{\lambda k}$	0.99	1.05	1.17	1.29	1.47

Fatigue	All raters	λ_k	-0.04	-0.04	-0.05	-0.05	-0.05
		$SE_{\lambda k}$	0.12	0.12	0.12	0.12	0.12
		$Infit_{\lambda k}$	0.98	0.99	1.00	1.01	1.02
		$Outfit_{\lambda k}$	0.99	1.01	1.02	1.04	1.04
	DRIFT	λ_k	-0.05	-0.04	-0.05	-0.05	-0.07
		$SE_{\lambda k}$	0.12	0.12	0.12	0.12	0.12
		$Infit_{\lambda k}$	0.98	1.05	1.11	1.25	1.33
		$Outfit_{\lambda k}$	0.99	1.08	1.16	1.33	1.44

Table 5

Accuracy of Standardized Difference Nominations for Primacy and Recency

Effect	True	Nomination	Effect Size			
			Small	Medium	Large	Very Large
Primacy	Non-Drift	Non-DRIFT	85.50	84.43	86.00	85.80
		DRIFT	4.50	5.57	4.00	4.20
	DRIFT	Non-DRIFT	5.83	0.60	0.00	0.00
		DRIFT	4.17	9.40	10.00	10.00
Recency	Non-Drift	Non-DRIFT	85.67	84.67	86.10	85.87
		DRIFT	4.33	5.33	3.90	4.13
	DRIFT	Non-DRIFT	5.67	0.80	0.00	0.00
		DRIFT	4.33	9.20	10.00	10.00

Table 6

Accuracy of Infit Nominations for Differential Centrality and Extremism

Effect	True	Nomination	Effect Size			
			Small	Medium	Large	Very Large
Centrality	Non-Drift	Non-DRIFT	87.43	87.20	87.70	87.03
		DRIFT	2.57	2.80	2.30	2.97
	DRIFT	Non-DRIFT	7.10	2.50	0.37	0.10
		DRIFT	2.90	7.50	9.63	9.90
Extremism	Non-Drift	Non-DRIFT	86.70	85.60	86.73	86.57
		DRIFT	3.30	4.40	3.27	3.43
	DRIFT	Non-DRIFT	2.97	0.30	0.03	0.00
		DRIFT	7.03	9.70	9.97	10.00

Table 7

Accuracy of Outfit Nominations for Differential Centrality and Extremism

Effect	True	Nomination	Effect Size			
			Small	Medium	Large	Very Large
Centrality	Non-Drift	Non-DRIFT	85.80	85.80	86.07	85.60
		DRIFT	4.20	4.20	3.93	4.40
	DRIFT	Non-DRIFT	6.60	2.23	0.33	0.07
		DRIFT	3.40	7.77	9.67	9.93
Extremism	Non-Drift	Non-DRIFT	84.90	84.20	85.30	85.33
		DRIFT	5.10	5.80	4.70	4.67
	DRIFT	Non-DRIFT	2.83	0.33	0.07	0.00
		DRIFT	7.17	9.67	9.93	10.00

Table 8

Accuracy of Variance Nominations for Differential Centrality and Extremism

Effect	True	Nomination	Effect Size			
			Small	Medium	Large	Very Large
Centrality	Non-Drift	Non-DRIFT	86.87	86.13	86.63	86.47
		DRIFT	3.13	3.87	3.37	3.53
	DRIFT	Non-DRIFT	7.27	3.87	1.17	0.50
		DRIFT	2.73	6.13	8.83	9.50
Extremism	Non-Drift	Non-DRIFT	86.13	85.97	85.53	85.90
		DRIFT	3.87	4.03	4.47	4.10
	DRIFT	Non-DRIFT	4.00	0.77	0.07	0.00
		DRIFT	6.00	9.23	9.93	10.00

Table 9

Accuracy of Infit Nominations for Practice and Fatigue

Effect	True	Nomination	Effect Size			
			Small	Medium	Large	Very Large
Practice	Non-Drift	Non-DRIFT	87.20	86.07	86.33	86.27
		DRIFT	2.80	3.93	3.67	3.73
	DRIFT	Non-DRIFT	9.43	8.67	7.30	4.90
		DRIFT	0.57	1.33	2.70	5.10
Fatigue	Non-Drift	Non-DRIFT	86.23	86.33	86.10	87.07
		DRIFT	3.77	3.67	3.90	2.93
	DRIFT	Non-DRIFT	9.10	8.70	6.93	5.23
		DRIFT	0.90	1.30	3.07	4.77

Table 10

Accuracy of Outfit Nominations for Practice and Fatigue

Effect	True	Nomination	Effect Size			
			Small	Medium	Large	Very Large
Practice	Non-Drift	Non-DRIFT	85.47	84.37	84.83	84.73
		DRIFT	4.53	5.63	5.17	5.27
	DRIFT	Non-DRIFT	9.17	7.87	6.33	3.70
		DRIFT	0.83	2.13	3.67	6.30
Fatigue	Non-Drift	Non-DRIFT	84.73	84.90	84.27	85.33
		DRIFT	5.27	5.10	5.73	4.67
	DRIFT	Non-DRIFT	8.67	8.07	5.77	4.37
		DRIFT	1.33	1.93	4.23	5.63